Assignments

Contents

This page will contain all the assignments you submit for the class.

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ``{r} ``` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Theo Athanitis.

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
#install.packages("datasets")
library(datasets)

#install.packages("knitr")
library(knitr) #used for knitting to a pdf
```

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```r
dat <- USArrests
```

It is useful to rename the dataset as it creates a copy of the dataset stored in this newly created variable that can now be modified without changing the original, such as adding a new column for states. Additionally, renaming it can make it easier to call/ reference in later functions.

Problem 2

Use this command to make the state names into a new variable called State.

```r
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```r
names(dat)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"     "state"
```

The variables contained in this dataset are Murder, Assault, UrbanPop, Rape, and state.

Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: Quantitative- the values of Murder are numerical values with measurment units as they record the amount of Murder arrests per 100,000 people in each state.

What R Type of variable is it?

Answer: The variable Murder itself is of the type character (as shown using typeof('Murder')), but the Murder values for each state are numeric doubles (as shown with dat[1,1])

```r
typeof('Murder')
```

```
## [1] "character"
```

```r
typeof(dat[1,1])
```

```
## [1] "double"
```

Problem 4

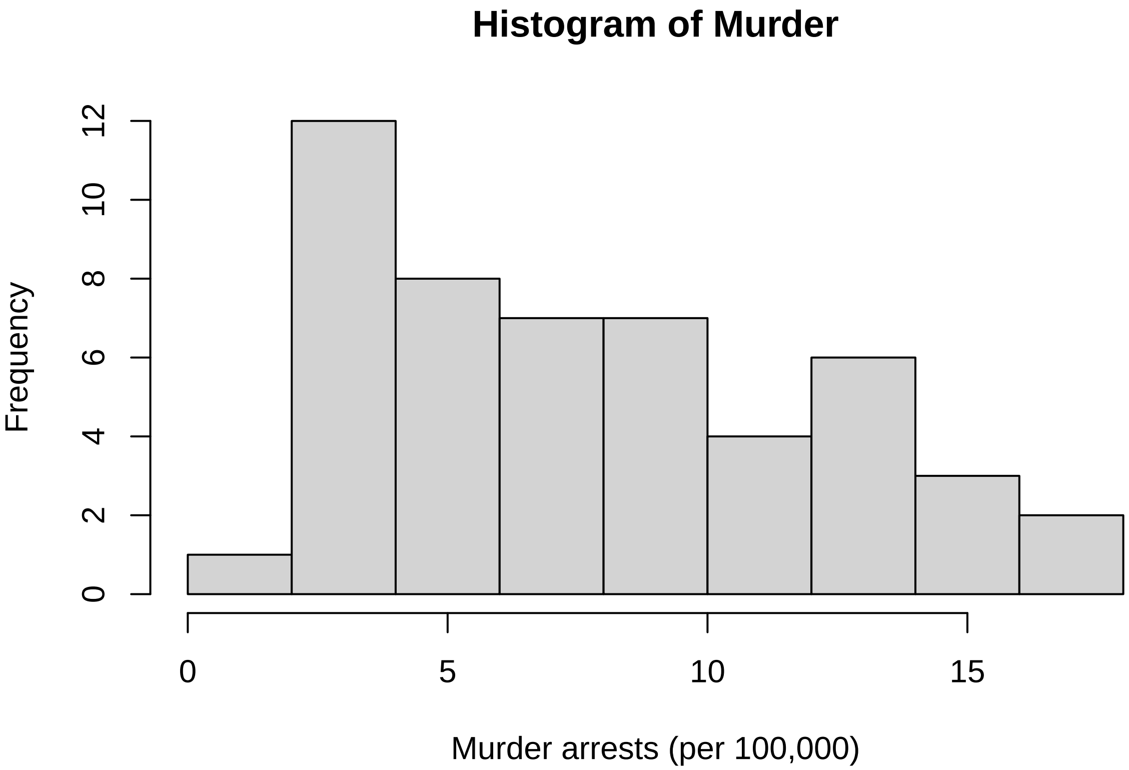
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This dataset contains data on violent crime rates and urban area populations for each of the 50 states in 1973. The variables and their values contained are 1. the number of murder arrests per 100,000 people, 2. the number of assault arrests per 100,000 people, 3. the percentage of the population that lives in an urban environment, and 4. the number of rape arrests per 100,000 people for every US state.

Problem 5

Draw a histogram of `Murder` with proper labels and title.

```r
hist(dat$Murder, main="Histogram of Murder", xlab="Murder arrests (per 100,000)", ylab="Frequency")
```



Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
mean(dat$Murder)
```

```
## [1] 7.788
```

```
median(dat$Murder)
```

```
## [1] 7.25
```

```
quantile(dat$Murder)
```

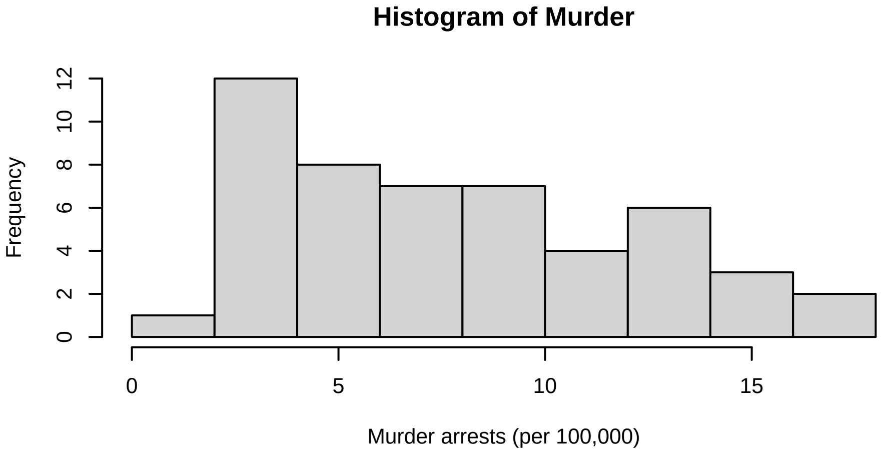
```
##     0%    25%    50%    75%   100%
##  0.800  4.075  7.250 11.250 17.400
```

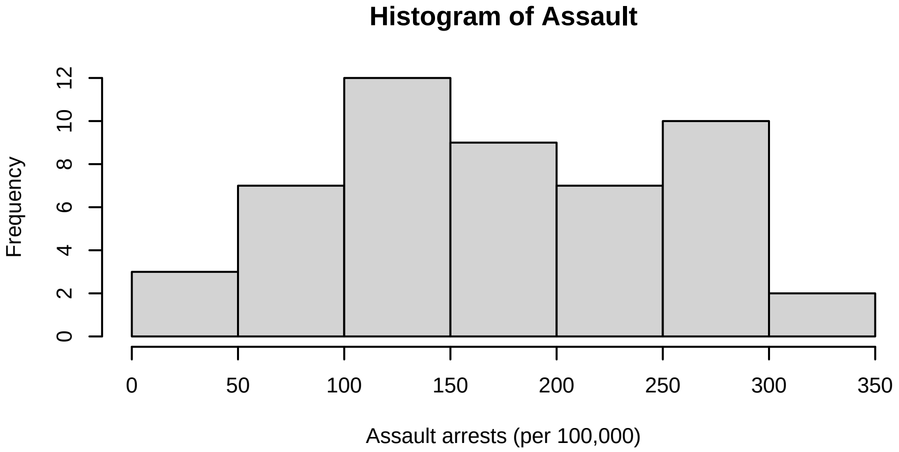
The mean of Murder is 7.788 and the median of Murder is 7.25. Mean is the sum of all of the states' murder arrests per 100,000 people divided by 50 (the number of states) also known as the average, while median is the middle value, where half of the value are greater than and half are less than the median, of the states' murder arrests per 100,000 people values. A quartile is one of three values which divide the dataset into four equal divisions. R likely only provides the 1st and 3rd quartiles (although I did not find a quartile function that operates this way and instead utilized the quantile function) as the 2nd quartile is the same as the median, therefore making the 1st and 3rd much more useful as they are more likely still unknown in comparison to the 2nd quartile when utilizing this function.

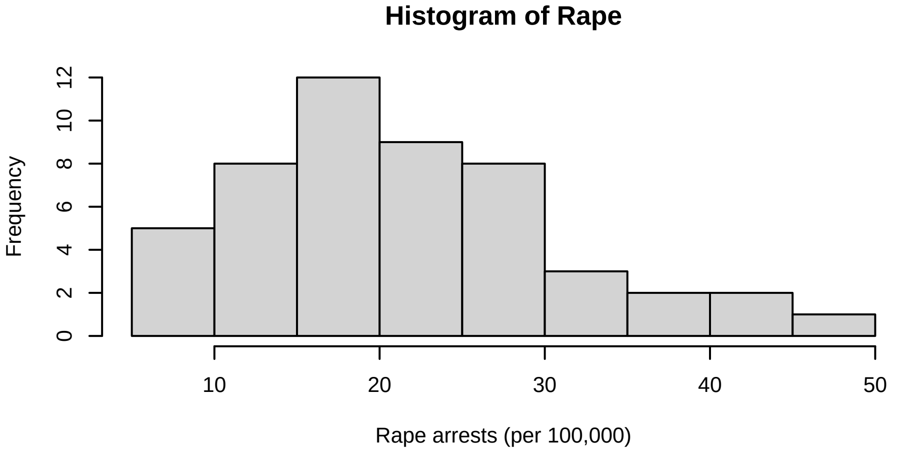
Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

```r
par(mfrow=c(3,1))
hist(dat$Murder, main="Histogram of Murder", xlab="Murder arrests (per 100,000)", ylab="Frequency")
hist(dat$Assault, main="Histogram of Assault", xlab="Assault arrests (per 100,000)", ylab="Frequency")
hist(dat$Rape, main="Histogram of Rape", xlab="Rape arrests (per 100,000)", ylab="Frequency")
```







What does the command par do, in your own words (you can look this up by asking R `?par`)?

Answer: The par command is used to modify the manner in which graphs are displayed by finding, modifying, or setting the parameters of graphs. One functionality of par includes the ability to show multiple graphs together in the same graphic as shown here. In this instance, the mfrow=c(3,1) parameter is a vector with subplots of 1 in length (row) and 3 in depth (column), to create the stacked graphs the function above produces.

What can you learn from plotting the histograms together?

Answer: By plotting the histograms together, it is easier to see the distributional differences between the different variables. By organizing the histograms this way, it can be seen that the frequency by state for murder and rape arrests per 100,000 people are skewed to the left, while the frequency of assult arrests by state are more evenly distributed. While these histograms are not directly comparable because of their different scales, skewdness can still be recognized across them and this display makes it easier to recognize.

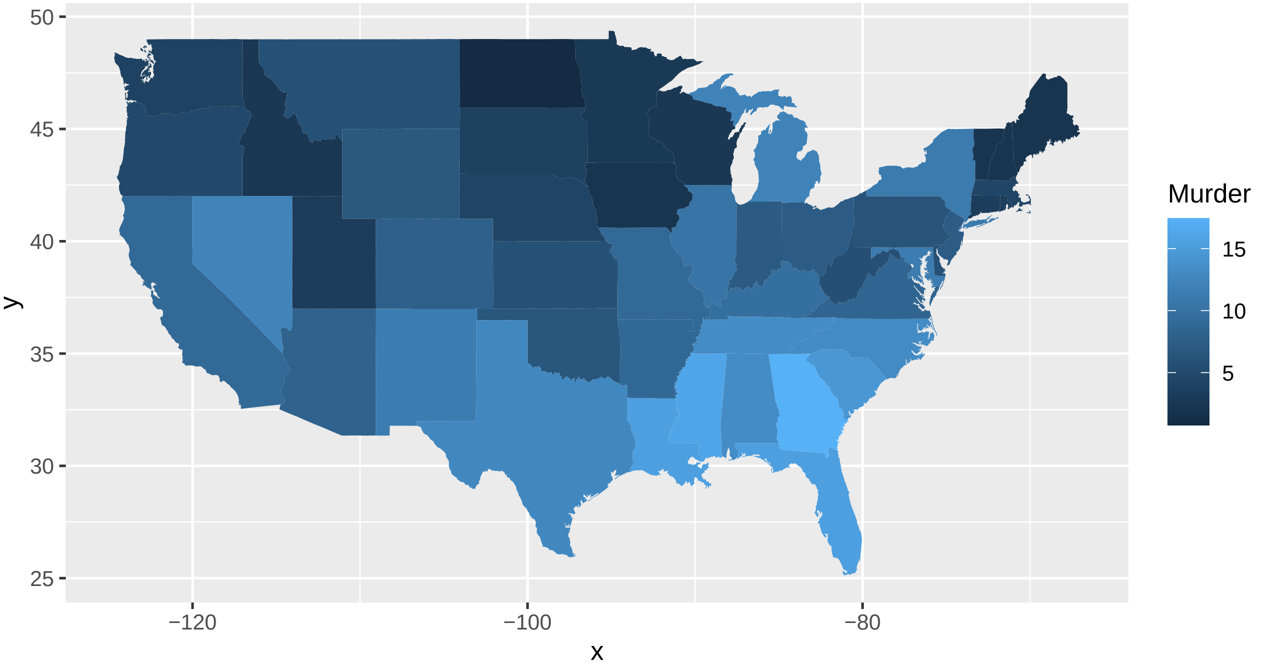
Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
library('maps')
library('ggplot2')


ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```



What does this code do? Explain what each line is doing.

Answer: First, the libraries for maps and ggplots are loaded. Next the ggplot function is called with first the parameter of dat as the dataset for the plot. The next set of parameters is for the aesthetic mapping for the plot basing the map_id for the values on the state variables and the fill/ color of that fill based on the Murder variable value for that representative map_id, as just defined. The fifth line further modifies the aesthetic mapping by defining the map for establishing the coordinate locations to display/divide the states and their fills. Lastly, the sixth line further modifies the aesthetic mapping by defining the x and y limits of this graphic based on the x and y limits from the state positional variables using their latitude and longitudinal values to ensure that all values are displayed.

These last three lines together are creating a map of the Murder arrests per 100,000 people for each state by scaling the color of each state on a map of the United States to represent the degree of this amount in comparion to the other US states.